Intersect Australia Ltd
PO Box H58, Australia Square
Sydney NSW 1215

T +61 2 8079 2590

ABN 67 131 752 657

NSW Coal Seam Gas

Data Background Paper

INTERSECT

Table of Contents

1 Executive Summary

Effective data management is an important element in achieving accurate oversight of Coal Seam Gas (CSG) activity in NSW. A correctly designed system will contribute to both real-time and long-term monitoring of environmental impact and resource management. Robust data management should enable provision of accurate information to government and the public, while ensuring that regulatory and compliance burden to industry is streamlined. Furthermore a well-designed system should ensure that information is both appropriately accessible and secure, and that privacy and confidentiality is maintained. Interfacing linkage and access protocols should also allow and promote broad access to, and use of, the data by government, public, industry and research bodies in order that impacts, risks and operations can be both managed and predicted with optimal accuracy.

At a primary level data management systems typically support data collection or ingest, secure storage, report generation and federated access. Systems typically seek to ensure data integrity and prevention of data loss, and to permit the re-use of data. Additional functionality can be incorporated to support different clearance levels, linkage across disparate datasets and advanced data analysis and modelling techniques. These substantially increase the value and utility of data collected.

The management of CSG data has specific characteristics requiring a design that provides the following functionality:

- Supports Commonwealth, State and Local Government laws, codes of practice and compliance systems;
- Provides differing access levels in a secure, robust environment;
- Retains interoperability across a wide range of custodians and stakeholders: policy and regulatory experts, monitoring and compliance bodies, natural resource managers, businesses, private land holders, the research sector, individuals and community groups;
- Uses open and widely supported specifications, standards and information management systems;
- Retains flexibility where possible, and is scalable, as information management systems and associated technologies are highly dynamic and the size of datasets is increasing at a rapid rate;
- Incorporates ability to ingest data of varying quality from diverse sources and therefore includes or interfaces with quality audit mechanisms;
- Enables electronic submission; and
- Enables linkage across diverse datasets.

The case studies contained in this paper draw out these key issues and how existing systems have attempted to address them. A high-level federated data management model that provides the flexibility required to deal with numerous data providers and a range of data types and formats whilst ensuring the necessary data security and integrity is outlined.

2 The Importance of Data Management for Coal Seam Gas

The design and operation of an efficient data management system is integral to the broader management and oversight of coal seam gas (CSG) activity in NSW. Effective data and information management will ensure that authoritative, reliable and up-to-date CSG related data and information is available to support government policy and decision-making, monitoring and compliance, business activity and transparent provision of information to the community. It will assist in achieving alignment with corresponding frameworks across NSW and other Australian jurisdictions. A comprehensive coordinated and defined approach is required to allow collection, submission, management of, and access to, a diverse array of information and data related assets that span organizational, geographic and temporal boundaries.

As the NSW and Federal Governments continue to develop policies governing CSG activity in NSW, mechanisms to provide coordinated submission, collation and access to reliable extensive and detailed data will be a key consideration. Furthermore current technologies for data storage, access to cloud-based high-performance computing, data mining, linkage, advanced analytics, machine leaning and modelling hold particular potential in ensuring that the CSG industry operates effectively and that the governing regulatory, monitoring and compliance environment is based on best available up-to-date information.

2.1 Government Framework

The functionality and specifications of an effective data management system will primarily be driven by the Federal, State and Local Government regulations, codes of practice, monitoring, compliance and reporting requirements that govern the operation of coal seam gas activities in NSW. Primarily these relate to environmental, ground and surface water quality and resource management, and workplace health and safety.

Primary responsibility for the regulation of the coal seam gas (CSG) industry in NSW rests with the Division of Resources and Energy in NSW Trade & Investment, the Environmental Protection Agency, NSW Office of Water, Department of Planning and Infrastructure and WorkCover NSW. There are numerous additional bodies that should have input to development of CSG data management functionality. These include but are not limited to:

Land and Property Information, Department of Finance and Services, National Parks and Wildlife Service, Office of Environment and Heritage, Department of Primary Industries, Department of Premier and Cabinet, Office of the Chief Scientist and Engineer, Health and Emergency Services and Local Councils. Relevant advisory functions include the NSW Water, Information, and Privacy Commissioners, the Natural Resources Commission, and Regional Development Authority. At a national level the Australian Government Department of Sustainability, Environment, Water, Populations and Communities, National Resources Commission and Geoscience Australia, in particular, would have strong relevance.

2.1.1 NSW Environment Protection Authority

The NSW EPA is the lead regulator for environmental and health impacts of industry in NSW. The EPA now occupies an expanded role in relation to CSG. Previously it regulated the production aspects of CSG, but now exploration and assessment activities are included within its remit[1]. The EPA serves as the lead regulator for environmental and health impacts[2] and can take a variety of regulatory measures from giving advice, providing education programs, and issuing warning notices, through to prosecutions. It issues Environment Protection Licences to all activities related to CSG.

The full text of individual licences, are available on the EPA's public register, along with statements of compliance, penalty notices, results of civil proceedings, and many other records relating to licences[3]. In addition to making these records available, it should also be noted that under "s320 of the [Protection of the Environment Operations Act 1997] application can be made to the EPA for access to monitoring data which has been submitted to the EPA by licensees"[4].

A good sense of the monitoring and reporting required of CSG operations is given by studying a relevant Environment Protection Licence[5]. This license requires the licensee to:

- Maintain a register of the gas gathering reticulation system, including gas well location, well head configuration and trunk lines (A2.4);
- Monitor air emissions (P1.1) for a range of chemicals and ensure they are within stipulated load limits (L2) and concentration limits (L3);
- Monitor ground water quality (P1.2);
- Monitor and ensure noise (in normal operation and during flaring events) is within specified limits as measured at specified locations (L5); and
- Monitor for the presence of potentially offensive odours (L7).

Results of monitoring must be recorded and retained for at least four years after the monitoring event and be able to be produced in legible form at the request of an authorised officer of the EPA (M1). Records must include:

- The date(s) the sample was taken;
- The time(s) at which the sample was collected;
- The point at which the sample was taken; and
- The name of the person who collected the sample (M1.3).

The monitoring frequency varies, but for some pollutants and locations it is continuous and for other pollutants and locations it can be quarterly or annually. A sampling or testing method is specified in each case (M3, M4).

---

[1] http://www.epa.nsw.gov.au/licensing/csgchanges.htm
[2] http://www.epa.nsw.gov.au/licensing/csgqanda.htm
[3] http://www.epa.nsw.gov.au/prpoeo/index.htm.
[4] http://www.epa.nsw.gov.au/prpoeoapp/ViewPOEOLicence.aspx?DOCID=32564&SYSUID=1&LICID=12003, http://www.austlii.edu.au/au/legis/nsw/consol_act/poteoa1997455/s320.html
[5] In the present report we have consulted licence 12003.

Monitoring activities are reported in a "Monitoring and Complaints Summary" as part of an Annual Return. The EPA supplies the licensee with a form that must be completed and returned by registered post. The licensee must retain a copy of the report for at least four years (R1).

The licensee must submit a Leak Detection and Repair Program Summary Report and a Ground Water Monitoring Report with the Annual Return. It must also submit updated spatial information when infrastructure changes have occurred (R4.5)[6].

2.1.2 NSW Office of Water

The role of the NSW Office of Water in overseeing coal seam gas activities is summarised by the Office of Coal Seam Gas as follows:

> *The NSW Office of Water is responsible for the management of the State's surface water and groundwater resources. The Office of Water will assess the potential impacts of a coal seam gas proposal on water resources, their dependent ecosystems, culturally significant sites and existing water users. This assessment will cover potential impacts on water table levels, water pressure, and water quality and will be provided to the appropriate consent authority. Any development that is approved will be required to hold water access licences for the water that is taken from any affected water source*[7].

A water access licence issued by the NSW Office of Water is required for any CSG activity using more than 3 mega litres of water per year[8].

The Office has produced an Aquifer Interference Policy[9] that applies to all activities that could result in interference with an aquifer, including CSG activities (2). The policy places certain requirements on the proponents of CSG activities to provide information that feeds into the assessment process. Those requirements that are most likely to result in the collection, generation, analysis and modelling of data include:

- Establishing baseline groundwater conditions including depth, quality and flow based on sampling existing bores and any new monitoring bores. (26)

And providing:

- Details of potential water level, quality or pressure drawdown impacts on nearby water users and groundwater dependent ecosystems;
- Details of potential for increased saline or contaminated water inflows to aquifers;
- Details of potential to cause or enhance hydraulic connection between aquifers; and
- Details of method for disposing of extracted water. (26)

And in the case of approved projects:

---

[6] This must be in ESRI geodatabase or shapefile format or any ESRI compatible dataset in the GDA94 coordinate system.
[7] http://www.csg.nsw.gov.au/protections/oversight-of-the-coal-seam-gas-industry#.UdUSsT5RRJV
[8] http://www.csg.nsw.gov.au/protections/aquifer-interference-policy#.UdUOGD5RRJU
[9] http://www.water.nsw.gov.au/ArticleDocuments/34/nsw_aquifer_interference_policy.pdf.aspx

- Details of an effective groundwater/surface water/pressure, flow and quality monitoring program through all phases of activity;
- Details of appropriate water measurement devices, regimes or methods such as water meters;
- Details of appropriate reporting procedures for results of monitoring and metering (26); and
- Estimates of quantities of water based on various levels of modelling or analysis (27).

2.2 Use of Data: Accessibility

Data management regimes need to provide broad and efficient access to data, incorporating mechanisms that ensure community confidence and transparency with robust systems for retaining privacy and commercial confidentiality where required. In addition to functionality required under the regulatory, compliance and monitoring environment referred to above, technical architecture will also be determined by current open government and open data policy initiatives that aim to increase transparency, integrity, and innovation. These are being deployed internationally, nationally and specifically in NSW.

In the case of access to CSG information it is important to note that "Open" implies the existence of identifiable, catalogued, quality, well-maintained datasets that are accessible according to defined access protocols that protect relevant confidentiality, privacy, and commercially sensitive details. Under such protocols full-level access may be limited to appropriate regulators while de-identification protocols may allow access to aggregated data for the purposes of ongoing long-term modelling and ongoing research.

In NSW, legislation supports the release of government data: the Government Information (Public Access) Act (2009) commenced on 1 July 2010, and provides a 'push' model where agencies are required to publish certain information and are encouraged to proactively release information. The NSW Government (as well as Federal and other State and Territory governments) is currently developing an Open Data Policy as part of the NSW ICT Strategy[10] and supporting technical architecture to make Government data available. Furthermore, a number of agencies including the NSW Department of Finance and Services, Land and Property Information, Office of Environment and Heritage (Open OEH) and Department of Transport are making significant progress in making data available.

10 http://www.finance.nsw.gov.au/ict/sites/default/files/NSW%20Government%20ICT%20Strategy%202012_1.pdf

Developing Open Data Policy In NSW

As part of the NSW ICT Strategy, the NSW Government Department of Finance and Services recently released the Open Data Policy Consultation draft (May 2013)[11]. The draft outlines the following ten principles for governing data release by NSW agencies[12]:

Open by default

Data is open by default unless there is an overriding reason for data not to be released in accordance with the Government Information (Public Access) Act 2009 (NSW) and the public interest test.

Protected where required

Data should not be released, or not released in full, where privacy, security, confidentiality or legal privilege considerations preclude its release. Information labels and security classifications can help indicate whether data is protected. Personal or identifying information may need to be removed from datasets, in line with the Government Information (Public Access) Act 2009 (NSW) and other applicable privacy policy and legislation.

Particular care, and in some cases further investigation, may be required where disparate datasets, individually de-identified, could potentially be linked or combined, to re-identify individuals, or breach relevant privacy legislation or policies.

Prioritised

High-value and requested datasets will be prioritised for release, in line with demand from the public and industry, or where the release of the datasets will contribute to better service delivery in NSW.

Discoverable

Data will be easily discoverable and searchable, with good metadata, and to further support this, will be published through a single, easy-to-use data portal: data.nsw.gov.au.

Usable

Data will be in a format that makes it easy to use, transform and reuse. Characteristics that support data usability include:

- Machine-processable formats;
- Non-proprietary formats;
- Completeness; and
- Clear, high quality metadata.

Primary

Data should be published as collected at the source, with a high level of granularity, and not in aggregate or modified forms. Data will be collected in an impartial and ethical way, in accordance with

[11] http://engage.haveyoursay.nsw.gov.au/document/show/933
[12] These principles have been taken directly from the consultation draft, and include relevant context as it appears in the draft. Not all context is included.

best practice collection methodologies, NSW and Australian standards (such as the ABS Data Quality Framework 2009), and applicable privacy legislation or policies.

Timely

Data will be current, and if practicable, live – with real-time feeds provided as appropriate.

Well managed, trusted and authoritative

Data will be well managed to ensure its ongoing integrity and efficacy for users. Appropriate data governance arrangements will be established, and data will be maintained in accordance with best practice information management principles. Applicable policies include the NSW Data and Information Custodianship Policy within the broader NSW Information Management Framework.

Free where appropriate

Data should be provided free of charge where appropriate, to encourage its widespread use for innovation, achieve the maximum value from the data for the citizens of NSW, and to enhance transparency of government.

Subject to public input

Data portals should be subject to public input, and have mechanisms by which users can engage with the data provider, and with the broader community of stakeholders around the dataset.

Mechanisms to ensure secure, appropriate levels of access to, and maintenance of, relevant privacy and confidential details have been developed and are in operation in NSW, particularly in management of health records and associated linkage and can be developed appropriately to allow access to CSG data. Examples are detailed further in Section 4.

Example: How conditional access to coal bed methane data is managed by the Alberta Energy Regulator in Canada: Confidential Well List

The Alberta Energy Regulator operates a record management system that is based on open access data principles and which allows public scrutiny of most CBM well records. However, certain wells are assigned a confidential status, following application for and subsequent classification as an exploratory well, or due to other defined extenuating circumstances defined in regulation as "conditions beyond the control of the operator, preventing the operator from utilizing the competitive advantage from the information obtained from a well." An accessible catalogue of confidential well information detailing well locations, licence numbers, licensee operator codes, company names, well confidential types, confidential depths for CB wells and the confidential release dates is maintained and updated daily by the Alberta Energy Regulator.

2.2.1 Industry Requirements

There are two primary issues that relate to CSG exploration and production: minimising compliance and reporting burden associated with compulsory reporting, and ensuring that commercial confidentiality is retained.

At a user level this implies that reporting is electronic and web-enabled. As numerous reporting requirements are likely to exist, where possible the data provided should be linked to avoid duplication. This would require architecture that enables linkage across different agencies and bodies, and which retains varying levels and permissions of access and associated protocols. Mechanisms with which to do this are detailed further in case studies described later in this document.

In addition to CSG operators it is likely that a broad range of industries would make use of the data. This includes local businesses, services, logistics, investment and finance sectors.

2.2.2 Ongoing Modelling: Role of the Research Sector

Establishment of long-term environmental baselines is an essential element of environmental impact monitoring, assessment and prediction. The research sector contributes at a foundational level to both the data-dependent knowledge-based management of environmental resources, and to development of measurement and analytical techniques. Research groups generate and hold significant datasets relevant both to real-, or near-time monitoring and construction of long-term baselines and predictive models. It is important that data management systems interface appropriately and provide input to CSG-related monitoring, management, data science, environmental management and extraction technologies.

There is extensive research and infrastructure capacity in universities, associated centres of Excellence, CRCs and Nationally funded research infrastructure facilities, National ICT Australia (NICTA), CSIRO, the Australian Nuclear Science and Technology Organisation, the National Measurement Institute, Geoscience Australia, and NSW State Government primary industries and environment departments:

- NICTA has strong expertise in developing advanced data analytics, optimisation, machine learning and modelling techniques. These are being applied to environmental and groundwater modelling using data from public, private and research sources. Machine learning and data fusion are specifically being applied to develop predictive groundwater models to assess potential risks associated with coal seam gas exploration, and energy exploration;
- ANSTO uses nuclear analysis techniques to analyse natural water systems, and inform water resource management and trace the sources of air pollution;
- CSIRO, through the Water for a Healthy Country Flagship, has developed significant groundwater quality, modelling and resource management expertise. The Earth Science and Resource Engineering division are developing CSG production techniques and forecasting methodologies as well as economic extraction and fugitive emissions monitoring and measurement systems. CSIRO also has a research program directed at understanding the complex, multi-scale mechanics of hydraulic fracturing at a fundamental level in order to develop numerical models, monitoring methods, and powerful and flexible hydraulic fracture technologies;
- eResearch infrastructure: Australia has a national eResearch system of high-end IT capacity that is designed to provide the technology, infrastructure and tools required to support increasingly data intensive research. This includes high-performance computing, scalable storage infrastructure, data management and discovery tools and a national research cloud that is developing on line mechanisms to process, link and use data. The Australian Access

Federation facilitates assignment of trusted identities to facilitate trusted communications and collaboration within and between institutions;

- National Measurement Institute (NMI) develops and provides measurement services and solutions for environmental impact assessments and regulatory compliance covering a range of chemicals applicable to coal seam gas (CSG) extraction. Laboratory-based water and chemical testing capabilities can be applied to baseline quality testing of groundwater and effluent or processed water prior to its re-injection, re-use or release back into the environment. NMI also performs air quality testing for volatile organic compounds and develops reference standards for other laboratories;
- Geoscience Australia (GA) operates geological observatories and monitoring systems, and is the peak body for development, maintenance and provision of national high quality geological datasets. This includes collections that support CSG exploration, monitoring and modelling.

3 The functions of a data management system

This section outlines the typical functions and considerations of a data management system and focuses on those aspects most relevant to kinds of data generated and required by the CSG industry.

3.1 Reporting, Compliance, Monitoring

Data management systems generally have some provision for reporting, although how extensive this is varies markedly. A system may, for instance, provide rudimentary search operations that permit data to be summarised by type, format and date range. In these cases, compliance and monitoring occur externally.

More sophisticated systems will incorporate compliance and monitoring workflows. As new data is deposited, for instance, such a system may alert administrators to its submission, so that it can be checked for compliance. An administrator acting on the alert can then step through the compliance checking process and record the outcome of each step within a tracking system. The Secure Unified Research Environment's Curated Gateway, in use in the NSW health and medical research community, is a case in point. Under SURE-enabled protocols, data cannot enter or leave the system without prior review and approval. The review workflow is effectively built into the data management system.

Systems supporting data management often need to support complex security models to handle authentication and authorised access to data products. CSG industry data is no exception, since a proportion of the data is publicly accessible while at the other end of the spectrum a range of datasets will be confidential and accessible to a very restricted number of users. Authentication and authorisation systems will also need to support a wide range of data consumers and modes of data deposit (including possibly automated forms of data capture).

3.2 Data Collection Mechanisms

Data management systems must incorporate ingest mechanisms for all the data formats they support. The key properties of each data type, such as its size, complexity, and mode of collection (be it automated or manual) will affect the approach taken to data deposit. Due to the heterogeneity of CSG related datasets it is likely to be necessary to support a number of modes of deposit. These modes could include web-based forms and bulk upload mechanisms. In addition attention must be paid both to how ingested data is best validated and the incorporation of appropriate supporting standards.

Online deposit has a number of advantages over manual submission of data collections including more efficient handling of data, reduced data re-entry, and often, faster submission times. However, it may place additional requirements on data submitters with respect to technical infrastructure (such as computing platforms and network bandwidth) and staff trained in the submission process.

3.3 Secure Storage

At a fundamental level, secure storage involves ensuring that all physical storage of data is on redundant storage, backed-up and versioned as appropriate. The objective is to ensure that at the physical level nothing is ever lost.

Additionally, secure storage entails that the privacy and confidentiality of data is maintained. In part, at least, this involves providing physical security of data centres and taking appropriate steps to prevent unauthorised electronic transfer of data, such as through the use of firewalls.

Lastly, secure storage requires data collections are annotated with sufficient metadata to ensure their future discoverability and reuse. Many important scientific datasets, such as climate records, are longitudinal in nature. This is true of a sizable number of datasets relating to CSG, particularly those relating to groundwater quality and flows and geological characteristics. Being able to study data relating to significant periods of time is one enabler of ongoing modelling and research. Accordingly, conditions of collection (such as unit sampling or real-time sensor network recorded data), post collection processing, and other details need to be recorded along with the data. Sufficient metadata is a data security issue, because without it, data kept in otherwise durable storage may nevertheless be effectively lost.

3.4 Evaluation, Analysis, Modelling, Linkage

At one level, data management systems are concerned with maintaining data integrity. They ensure that data is securely stored, that there are clearly defined mechanisms for its deposit, and that it can be discovered, recalled and reused in the future by those with authority to do so.

The last aspect allows a data management system to enable additional activities, such as evaluation, analysis, modelling and data linkage, either by pushing data to other systems or by internally supporting those activities. Interfaces to push data onto compute resources for intensive analysis is often a desirable feature of a data management system. Meanwhile, data linkage (the joining of separate datasets into one larger dataset) has been proven to yield new insights, as evidenced for health data in the Secure Unified Research Environment case study in Section 4.2.

3.5 Complexity relating to CSG data

Data associated with CSG production is complex across several dimensions. It is produced by many parties with differing primary interests, e.g. different State and Commonwealth Government departments, commercial entities and Non-Government Organisations (NGOs); is useful for multiple purposes (Monitoring, Decision Making, Regulation, Policy Development, Commercial Planning, Community Information, Research); and is of interest to a wide and varied audience (e.g. Government, Commercial, NGOs, Researchers and the Public).

3.5.1 Range of formats and information types

CSG-associated data covers a broad spectrum, and ranges from: geological data indicating the potential occurrence of CSG and other related systems (e.g. groundwater); to data concerning the particulars of the production of gas in a specific location and at a particular time; and data that may inform on any potential social and environmental impacts. Data may be static, real-time or historic.

Site-associated data pertains to where CSG is extracted and also other locations such as groundwater-associated sites (bores or springs) and the logistics networks required to collect and transport the gas to market. Site-associated data and metadata formats range from spatial to categorical and numerical information, to supporting documentation such as photographs, and includes:

- Site location and identification (spatial location, names etc);

- Geology and hydrology data;
- Details and specifications of equipment (CSG or bore water extraction equipment);
- Land use;
- Water monitoring information (groundwater levels, chemical and biological quality analyses); and
- Other environmental monitoring information (e.g. air quality & noise analyses).

Party-related data relates to organisations or individuals associated with the above sites, or who generate data associated with the above sites. It consists primarily of text-based information covering:

- Producers of geological, hydrological and environmental survey data;
- Petroleum tenure holders;
- Land holders;
- Individuals undertaking site monitoring and assessments (e.g. identification, qualifications); and
- Data modellers.

Temporal (date and time) information is also included in both site-associated and party-related data to record the monitoring regime employed and any differences across time.

3.5.2 Multiple data ownership

As touched on above, data relevant to the CSG industry is *produced* by many different parties; including multiple Commonwealth and NSW Government departments; commercial entities; research organisations and non-government organisations (NGOs).

As such, the data is *owned* by multiple parties who have specific business reasons for its production (planning, monitoring, compliance) and specific intended audiences (private, collaborative, public).

The use of data licensing frameworks (such as Creative Commons or the Australian Government's Open Access and Licensing Framework) by data owners clearly articulates the circumstances under which each data type may be accessed and for what purpose, as well as specifying attribution. As noted in the NSW Open Data Policy Consultation draft (May 2013) "Clear licensing frameworks will help to prevent the misuse of datasets, while also providing a high degree of flexibility for the licensee in their use of the data".

3.6 Data Integrity and Accuracy

It is essential to ensure the integrity and accuracy of data when it is used as the basis to inform and provide evidence for any decision or communication message. In order to achieve this goal, data must be appropriately managed through its lifecycle, both from a technology and human perspective.

From a technology perspective, any CSG data management system should minimise the chances of compromising data during normal procedures such as data entry, upload or data transfer via the use of data entry methods that prevent the input of invalid data, and the use of automated error detection

and validation mechanisms where appropriate. Additionally, adoption of a regular backup regime can be used to minimise the impact of events such as hardware malfunctions.

From the human perspective, business rules around curation of data in the system (e.g. data creation or receipt in the system, definitions, linkage, lineage through the system, etc.) are also required.

3.7 Confidentiality and Privacy (commercial and public)

Whatever the motivation for an organisation sharing their data in this arena (e.g. providing transparency, through to legal compliance responsibilities), any CSG data management system for NSW will need to address individual privacy and commercially sensitive information in accordance with appropriate legislation (e.g. NSW Privacy and Personal Information Protection Act 1998; Commonwealth Privacy Act 1988), and also maintain commercial confidentiality where appropriate. Data management models for ensuring privacy are available and may be adapted from mechanisms currently in use in the Health sector in NSW. These are described in detail Section 4.

4 Case studies of existing Data Management projects

The following case studies explore existing data management systems. Some of these systems have as their sole or partial function, the collection and management of coal seam gas data. The remaining case studies focus on systems that have had to cater to challenging data management issues such as supporting data linkage in the context of health records, where the preservation of privacy of paramount. It is noteworthy that the two health related mechanisms described can interface to allow secure access to multiple, diverse and distributed datasets and corresponding linkage. This expands significantly both the spectrum of users and analyses that can be carried out, with potential to increase impact in policy development, evaluation, business use and research.

4.1 Energy Resources Conservation Board, Alberta, Canada

Relevance to Coal Seam Gas industry data: Alberta is Canada's only province to have substantially exploited its CBM (Coal Bed Methane) resources[13]. The Digital Data Submission System supports data reporting under the Alberta regulatory and compliance system for CBM production. The system operated by the Alberta ERCB is substantial and operates on open access principles.

The Digital Data Submission (DDS) Service of the Alberta Energy Regulator (AER) "provides a gateway for customers to access and provide electronic information and data to the systems that the AER utilizes in conducting its business".[14] All coal bed methane (CBM) wells are licensed requiring that "data on CBM wells, including produced fluid volumes and rates (oil, gas, and water) must be collected and reported to the ERCB in the same manner as occurs for other gas wells in the province".[15] Additionally, produced fluids must be sampled, analysed and associated data reported. Data on experimental test wells must also be reported to the ERCB. The DDS allows this information to be submitted online.

The public-facing components of the online submission service[16] and the associated directives[17], suggests the regulator faces a number of data management issues:

- Handling many different data reporting products to provide comprehensive online submission of data required for its regulatory activities, including:
    - Application submission and licence management;
    - Audit data;
    - Field surveillance notifications, incidents, inspections and compliance monitoring;

[13] http://www.ogj.com/articles/print/volume-104/issue-28/drilling-production/coalbed-methane-expands-in-canada.html
[14] https://www3.eub.gov.ab.ca/Eub/Dds/anonymous/Logon.aspx?ReturnUrl=%2fEub%2fDds%2fDefault.aspx
[15] http://www.energy.alberta.ca/NaturalGas/753.asp
[16] https://www3.eub.gov.ab.ca/Eub/Dds/anonymous/Logon.aspx?ReturnUrl=%2fEub%2fDds%2fDefault.aspx
[17] http://www.aer.ca/rules-and-regulations/directives

- Incidents including casing failures and surface casing vent flow/gas migration;
- Reports relating to groundwater protection and gas plant sulphur balance;
- Gas removal data;
- Liability cost estimates;
- Oil field waste reporting;
- Well drilling and completion data;
- Well packer test results; and
- Well test data.

- Complex security model allowing for both public availability as well as limited access and submission by registered customers and the AER. The user guide[18] suggests that responsibility and access for different data reporting roles can be delegated to sub-units of the customer or other parties; and
- Confidentiality requirements and embargo periods since some of the data provided to the service or accessed from it has commercial value.

The AER's Digital Data Submission service is delivered through a web interface with a hierarchy of forms to handle collection of the above datasets. The system requires a username and password to access most parts of the system. As noted above, registered customers can delegate access if needed.

The system handles a broad range of datasets, including geospatial data (concerning, for instance, the spacing of wells). In some cases, such as for well completion data, data is submitted in a strict XML format following a documented schema[19]. In other cases, it appears to be entered directly on web (HTML) forms.

A brief investigation suggests the system exhibits some or all of the following issues:

- Complexity – The regulatory framework requires the submission of a sizeable number of datasets, leading to a complex interface.
- Recent downtime – The system appears to have recently experienced a period of significant downtime owing to hardware failure of some of its servers. The consequences of the downtime were significant, with delays incurred by staff and customers having to fall back to hardcopy forms[20].
- Browser incompatibility – The service supports Internet Explorer 6.0 and 7.0[21]. Other web browsers are not officially supported and new versions of Internet Explorer must be run in compatibility mode.

---

[18] https://www3.eub.gov.ab.ca/Eub/Dds/Help/DDSGuide.pdf
[19] https://www3.eub.gov.ab.ca/eub/dds/anonymous/apps/wde/Filespecs.htm
[20] http://www.aer.ca/rules-and-regulations/bulletins/bulletin-2013-11, http://www2.canada.com/calgaryherald/news/calgarybusiness/story.html?id=5697a812-7f1b-48e7-8224-62f562d41611
[21] https://www3.eub.gov.ab.ca/eub/dds/anonymous/BrowserRequirements.html

4.2 Secure Unified Research Environment (SURE)

Relevance to Coal Seam Gas data:

- Provides a stringent, privacy-assured environment for pre-approved multi party access and analysis of multiple de-centralised datasets;
- Allows analysis that leads to high impact data use;
- Provides a digital archive for long-term preservation of datasets;
- May be adapted for CSG data management.

The Secure Unified Research Environment (SURE) is a "secure computing environment that Australian researchers can log in to remotely in order to analyse health data from different sources such as hospitals, general practice and cancer registries"[22]. The system was developed and is operated by the Sax Institute, University of Technology, Sydney as part of the Population Health Research Network. It addresses a number of significant health data use, access and management issues:

- Privacy – The privacy of individuals is paramount. Those handling health data must take extreme care to ensure sensitive health records are not leaked and cannot be traced back to the individuals they pertain to, through reconstruction of linked data information.
- Data linkage – By linking health datasets researchers can address previously unanswerable questions. Researchers, for instance, have been able glean new insights into consistency of care by linking Medicare data with the Sax Institute's "45 & Up" study and data from the Pharmaceutical Benefits Scheme[23].
- Collaboration – Drawing upon the collective expertise of different groups has obvious advantages, however the strict requirements with respect to security can make collaboration challenging.
- Tight access controls – Data pertaining to different projects needs to be kept separate and restricted to project stakeholders. The movement of data between systems needs to be carefully tracked.
- Retention and Curation – Some health datasets consist of longitudinal studies that need to be retained across technology refreshes. In many cases health datasets need to be maintained for long periods, as recommended in the Australian Code for the Responsible Conduct of Research.

The SURE addresses these data management issues by providing the following services:

- Centralised infrastructure – All SURE infrastructure is located centrally in a secure data centre. Access to the system is mediated through multiple firewalls. Because all analysis of the data is done on this central infrastructure, it is easier to maintain security of the data;

[22] https://www.sure.org.au/site-resources/sure-fact-sheet-july-2012
[23] http://www.zdnet.com/big-data-a-sure-thing-for-healthcare-7000000291/

- Virtualised workstations – Rather than working on local desktop environments, researchers use a Citrix client to login to completely virtualised workstations hosted on the centralised SURE infrastructure. The virtualised workstations provide a full desktop experience with a suite of analytical tools to complete analysis of the data. As well as optimising the use of the underlying infrastructure, virtualisation means that projects can be conducted in their own sandboxed environment, ensuring that data acquired for a specific research project is used for that purpose alone. The virtualised workstations also permit collaboration allowing multiple approved stakeholders to logon to the same workstation, access and analyse the data;
- Two-factor authentication – In addition to password and certificate controls, a physical security token, a specialised USB key, is required to complete authentication;
- The Curated Gateway – All data transferred into SURE and research outputs exported from the system pass through the Curated Gateway. This gatekeeper ensures that all transfers are reviewed and assessed before they are allowed to proceed. The data linkage itself is achieved using existing, independent services, such as the Centre for Health Record Linkage (CHeReL), described in Section 4.3; and
- Digital archive – The archive provides encrypted storage of all datasets for the period of retention required, following which they are destroyed.

The solution has some issues, most of which are a consequence of the complexity derived from necessary security requirements:

- Two-factor authentication is more complex than simple password authentication systems, potentially requiring some user training;
- Citrix client software is required on the researcher's computer;
- Analysis requires a good internet connection;
- Bespoke software for remote authenticated analysis may be complex to install and require specialist maintenance.

4.3 The Centre for Health Record Linkage (CHeReL)

Relevance to Coal Seam Gas industry data:

- Provides a system that links records, data and associated metadata mappings from disparate sources;
- Maintains privacy;
- May be adapted for CSG data management.

The Centre for Health Record Linkage (CHeReL)[24] provides a health record linkage service to approved researchers, health planners and policy makers. It operates under strict ethics and privacy governance

[24] http://www.cherel.org.au/

to ensure patient privacy. Input data consists of over 70 million health-related records from at least 18 distinct and separate patient-related data sources across NSW and the ACT [25]. These include hospital admission, emergency department data, the NSW cancer registry, and notifiable conditions information, each of which contain a discrete and finite amount of demographic and clinical information. Matching algorithms are used to link records and bring together information that relates to the same individual, family, gender, age-group, place or event etc. from the different data sources. This allows health researchers and planners to generate rich datasets about certain cohorts or groups from the multiple input data sources.

Data management issues faced include:

- Ethics – All research concerning human participants requires approval by a Human Research Ethics Committee;
- Privacy & De-identification – Preserving the privacy of individuals is of central importance to the CHeReL, while allowing aspects of their clinical records to be shared;
- Variety of data sources – Over 18 core input datasets and several other external datasets are used as data sources for the CHeReL;
- Security – Since data sent to the CHeReL by a data custodian contains personal information, secure transfer mechanisms are paramount; and
- Confidentiality – Signed agreements are required to maintain confidentiality.

CHeReL has solved or mitigated theses issues as follows:

- Ethics – An ethics application must be lodged with CHeReL prior to the commencement of any data linkage activities. Studies linking NSW Department of Health data must seek approval from the NSW Population and Health Service Research Ethics Committee[26], which requires the submission of a National Ethics Application Form[27] and a completed Research Protocol[28].
- Privacy & De-identification – CHeReL complies with best practice in privacy preserving record linkage adopting the Kelman, Bass and Holman approach[29], which guarantees privacy protection. In short, the record linking and de-identification process[30] operates whereby:
    - CHeReL staff performing the linkage use demographic variables about individuals (e.g. age, name, gender etc) but do not have access to any clinical information (e.g. diagnosis, length of hospital stay etc);
    - Custodians of the input data collections only have access to data within their data collections; and

[25] http://www.cherel.org.au/master-linkage-key
[26] http://www.cancerinstitute.org.au/research-grants-and-funding/ethics
[27] https://ethicsform.org/au/SignIn.aspx
[28] http://www.cancerinstitute.org.au/media/129193/2011-07-research-protocol-template.doc
[29] Kelman CW, Bass AJ & Holman CD. 2002 Research use of linked health data – a best practice protocol. Aust NZ J Public Health 26(3):251-5
[30] http://www.cherel.org.au/how-record-linkage-works

    - Researchers receive only clinical data, which contains no identifying variables, or variables which provide a link back to the CHeReL "Master Linkage Key" (which provides a pointer to records for a person in the different input datasets).

- Variety of data sources – The core set of input datasets into the CHeReL system includes at least 18 distinct and separate patient-related data sources from NSW and ACT[31]. Additionally, CHeReL regularly links several other relevant health datasets across NSW[32] and can also link to other data sources on a case-by-case basis. Linkage between the disparate input datasets requires:
    - Identification and mapping of equivalent fields between different input data sources. This is enabled by the provision of a "Data Dictionary" by CHeReL across the input data sources, which includes the list of variables and metadata definitions used in each input data source; and
    - Software to automate record matching across different data sources. This is required as subtle differences in information may be entered into different databases (e.g. "Robert Plant" vs. "Bob Plant"). CHeReL currently uses Choicemaker[33], which is an open source record matching software package, which incorporates an automated blocking algorithm and aspects of machine learning to assign weights to each record.
- Security – It is recommended that custodians of input datasets to CHeReL perform an encryption step to generate project specific identifiers so that original record IDs or patient IDs are not sent to CHeReL. Additionally, as data sent to CHeReL does contain personal information, all files sent should be encrypted using WinZip or GPG software using a minimum of 128-bit AES encryption security. Additionally, to circumvent security issues associated with emailing data, CHeReL provides and requires the use of a secure file upload facility to upload encrypted data from the input data sources;
- Confidentiality – Before any data is released, the NSW Ministry of Health requests that a confidentiality agreement is signed by the researcher. Custodians of other collections should also obtain a signed confidentiality agreement and follow the data disclosure policies of their own organisation;

The solution appears to be fit for purpose as it has resulted in the completion of 147 linkage projects (2007-present)[34] and the publication of 115 research papers and 8 reports (2008-present) based on the combined data[35].

[31] http://www.cherel.org.au/master-linkage-key
[32] http://www.cherel.org.au/external-datasets
[33] http://oscmt.sourceforge.net/
[34] http://www.cherel.org.au/completed-projects
[35] http://www.cherel.org.au/publications

4.4 NSW Office of Environment and Heritage – NSW MER Strategy (Monitoring, Evaluation and Reporting)

Relevance to Coal Seam Gas industry data: Overall, the MER system is highly relevant to Coal Seam Gas data management given similar environmental reporting objectives, some overlap in dataset types, and a strong orientation to geospatial datasets.

The Natural Resources Monitoring, Evaluation and Reporting (MER) program of the NSW Office of Environment and Heritage (OEH) monitors and reports on the condition of natural resources and the pressures placed on them. MER is focused on improving information flow between local, regional, state and national data managers to support natural resource management decisions in NSW. It has a cyclical reporting pattern and depends on longitudinal datasets. It has twelve themes, including: Native Vegetation, Fauna, Threatened Species, Wetlands, Estuaries & Coastal Lakes, Land Capability and Soil. Other government bodies including the Department of Primary Industries and Natural Resources Commission are responsible for delivering elements of MER.

MER's key data management challenges are:

- Heterogeneity of datasets – MER data covers a wider range and includes extensive LIDAR imagery, spatial data, bathymetric records, species counts, photography, and significant numbers of derived datasets, databases and data products;
- Long-term storage – Many datasets are irreplaceable and have significant value beyond an MER reporting cycle;
- Large datasets – Imagery data holdings, for example, are in the petabyte region;
- Distributed workforce – OEH activities take place at many offices and locations throughout the state. A significant amount of data is also collected in the field; and
- Promoting sharing and reuse – The datasets of one theme are potentially of use to other scientists within the organisation but their existence is not widely known.

OEH has approached data management differently for spatial and non-spatial datasets. Spatial data is data that is predominantly understood and interpreted in geographic terms. Non-spatial datasets may still have a geographic attribute (for instance, the name of a catchment to which the data points relate) but are interpreted more readily in terms of readings, counts, etc.

For spatial data, which constitutes approximately 80% of data assets by size, OEH has provided a petabyte-scale hierarchical storage management (HSM) facility. All data stored in the facility is carefully processed before deposit to ensure it meets a strict set of file naming conventions and directory structures. These procedures and standards ensure that the required data can be efficiently and accurately recalled as required.

For non-spatial datasets, which are highly heterogeneous, OEH has taken a number of measures:

- Expanded corporate storage – Increasing the amount of centrally managed storage to discourage the use of portable media;

- Shared storage consolidation – Structuring shared storage more stringently into spaces to store datasets for each of the MER reporting themes;
- Data management training – Making training available to scientists in data management best practice, such as following file naming conventions, metadata and creating data documentation;
- Metadata statements – Entrenching the practice of creating metadata statements describing data products;
- Registration and discovery services – Providing light-weight project data registration and discovery processes and systems to promote sharing and reuse of datasets; and
- Folder structures and file naming conventions: Providing tools to assist with the creation of documented folder structures and naming conventions for projects.

The "spatial solution" provides a highly structured, highly codified approach to managing large amounts of geospatial data. Because the data is effectively curated, data deposit requires the assistance of a technician - it is not a self-service tool for scientists.

The "non-spatial solution" implemented is the result of extensive consultation with scientists in a representative subset of the MER themes. The solution is guided by a desire not to displace the work practices of the scientists unnecessarily. It also provides, at some level, for data management in the field. For this reason, it supports a predominantly file system centric approach to data management. This limits to a degree the amount of automation, validation and enforcing of underlying business rules that can be done.

4.5 NSW Land & Property Information – NSW Foundation Spatial Data Framework

Relevance to Coal Seam Gas industry data: Spatial data is an important part of NSW data infrastructure and is implicit in a wide range of management, assessment and reporting activity and particularly for the management of CSG activity. The Foundation Spatial Data Framework when finalized will set out data management system highly relevant to CSG

NSW Land and Property Information, is in final stages of developing and releasing a NSW Foundation Spatial Data Framework (FSDF) that sets out a system for managing and accessing "fundamentally significant and important spatial data critical to the functioning, planning, maintenance and management of the State of NSW". Furthermore: "The emphasis of the NSW FSDF is to enable the seamless exchange of single source of truth spatial data, ensuring that the integral function and expectation of the community to the Government is maintained to the highest possible level of integrity in the delivery of a critical service".

Within this, to be classified as "Foundation" data must be considered:

- Essential for public safety and wellbeing;
- Critical for a national and state or government function;
- Contribute significantly to economic, social and environmental sustainability;

- Enable innovation by government, industry, research and academic sectors

The NSW Foundation Spatial Data Framework consists of the following ten foundation spatial data themes that have been identified through consultation and analysis of user requirements:

- Geocoded Addressing;
- Administrative Boundaries;
- Positioning;
- Place Names;
- Land Parcel and Property;
- Imagery;
- Transport;
- Water;
- Elevation and Depth; and
- Land Cover

The FSDF when finalised will identify an assembly of state-wide foundation spatial data that is important to a range of users including Federal, State and Local Government agencies, community, universities and private businesses. It will consist of spatial data themes and related datasets, to be managed under policy and governance structures that account for custodianship, deployment of standards, metadata, access, pricing, intellectual property and licensing, and which maintain appropriate privacy and security. Significant effort has been made to ensure alignment with similar frameworks across Australia.

The framework is supported by an extensive set of custodianship guidelines and associated standards that support interoperability and compatibility in terms of location specific data characteristics that include format, reference system, projection, resolution and quality. The guidelines set out the rights and responsibilities associated with capturing and managing information on behalf of NSW Government. They infer that custodianship and associated responsibilities for maintenance and data availability will be assigned to NSW Government agencies and organisations including State-Owned Corporations, Public Trading Entities, other State Government funded organisations and Local Government authorities. In addition to custodian responsibilities, the guidelines identify at a general level the obligations of data producers, distributors, users and aggregators. The Land and Property Information, Location Policy and Coordination Unit (LPCU) will also manage, maintain and provide access to a register of NSW spatial custodians and of the datasets for which the custodians are responsible.

Due both to the high relevance of location data to CSG activity, and the data management frameworks in place and under development, further analysis of, and integration with, location-based data management systems would be warranted.

4.6 Queensland Department of Natural Resources and Mines

Relevance to Coal Seam Gas industry data: This system is used in monitoring the potential effects of the CSG industry on groundwater in QLD. It is highly relevant for CSG data management in NSW.

In Queensland, the major areas of CSG production (the Surat and Southern Bowen Basins) overlie the Great Artesian Basin, which is an important source of groundwater for pastoral activities. Any potential impacts of CSG mining on groundwater resources requires a thorough understanding of groundwater systems and water flows, and monitoring of groundwater quality.

Under the Queensland Water Act 2000, petroleum tenure operators undertaking CSG extraction are required to carry out baseline assessment of water bores before commencing production, and to "make good impairment of bore supplies now and into the future".

In areas of concentrated CSG development in Queensland, impacts on water levels caused by individual CSG projects can overlap. This is referred to as a 'cumulative management area' (CMA). When a CMA is declared, the Office of Groundwater Impact Assessment (OGIA) in Queensland is required to prepare a 'cumulative assessment of impacts of CSG water extraction', develop integrated regional management arrangements, and set these out in an 'Underground Water Impact Report' (UWIR).

As part of generating the impact report, a regional groundwater flow model is developed which can be used to predict future water level impacts in coal seams as well as in adjacent aquifers. The report lists 'Immediately Affected Areas' and 'Long Term Affected Areas', and the petroleum tenure holders associated with these sites that are required to carry out ongoing baseline assessments.

Many types of data, are collected for each baseline assessment, at each bore, by various Petroleum Tenure Holders, including:

- Data quality measures, including accreditation and qualifications of the persons conducting the assessments, standards, quality assurance, requirement of independent third party certification;
- Person and organisation details);
- Bore identification information (including names and GIS data);
- Bore Information - construction, pumping equipment, water supply type (e.g. stock watering, domestic use);
- Standing water level measurement;
- Water quality; and
- Supporting documentation (eg. photos of the bore, other documents (e.g. water quality laboratory report, bore use logs, landholder agreements, historical records from the bore, etc)).

Data is uploaded from multiple tenure holders, compiled and managed at OGIA. Following this a regional groundwater flow model is developed for each CMA, and applied to the observed data and water level predictions made across the region. These predictions are included in the impact report (including representation on maps and specific information about bores that are predicted to be affected) which is to be updated every 3 years. Bore owners in the CMA can obtain specific information about the predicted impacts on their bore(s) using an online search mechanism[36].

Specific and comprehensive guidelines for best practice in the data collection are set out in the Baseline Assessments Guideline Document published by the QLD Department of Environment and Heritage Protection. Data is entered by tenure holders using a word document containing macros. Data entry is required to follow the definitions outlined in the Bore baseline assessment database data dictionary and Bore baseline assessment database data file format documents.

Documents can be sent in one of 2 ways:

- For tenure holders with fewer than 20 Baseline assessments, the documents can be scanned and emailed, or sent by post, to OGIA. Supporting documentation must conform to File Naming Standards (outlined in the Guidelines document).
- For tenure holders with more than 20 Baseline assessments, the information is expected to be sent in electronic format. The zipped files must be copied to a CD or DVD and posted to OGIA.

A limited search ability is available: Bore owners in the CMA can use an online search tool to search for specific information about predicted impacts on their bore(s).One search term is supported - the bore's registration number. e.g. search using "12340" for a bore in an IAA).

While this is a simple primarily manual system of data capture, the use of word documents, and the option of hard copies (and the allowing of 'electronic' submission of zipped data by larger operators), allows for universal coverage. The system has been successfully used to collect, collate and analyse data that has been suitable for production of the Surat UWIR.

[36] http://dnrm.qld.gov.au/ogia/surat-underground-water-impact-report/bore-search

5 Coal Seam Gas Data Management – A Federated Approach

As should be evident the collection and management of data relating to CSG activities involves many operators and data formats. One method for handling this diversity is to adopt a federated approach to data management. In a federated data management system individual groups are responsible for collecting and managing their own data, the federation describes the way in which the data held by these entities can be discovered and shared.

There are two main types of data federations. The first aggregates information on available data (metadata) and allows the datasets to be fetched in their original format. The second also aggregates the metadata on available datasets but in addition it provides a number of adapters to convert data from multiple disparate data sources into a single homogenous result set. This second approach is sometimes known as data virtualisation. Data virtualisation approaches usually have significantly higher implementation costs and complexity due to the need to develop adapters for each type of data present in the federation. However, this provides significantly enhanced ease of use for the end-users of the federation.

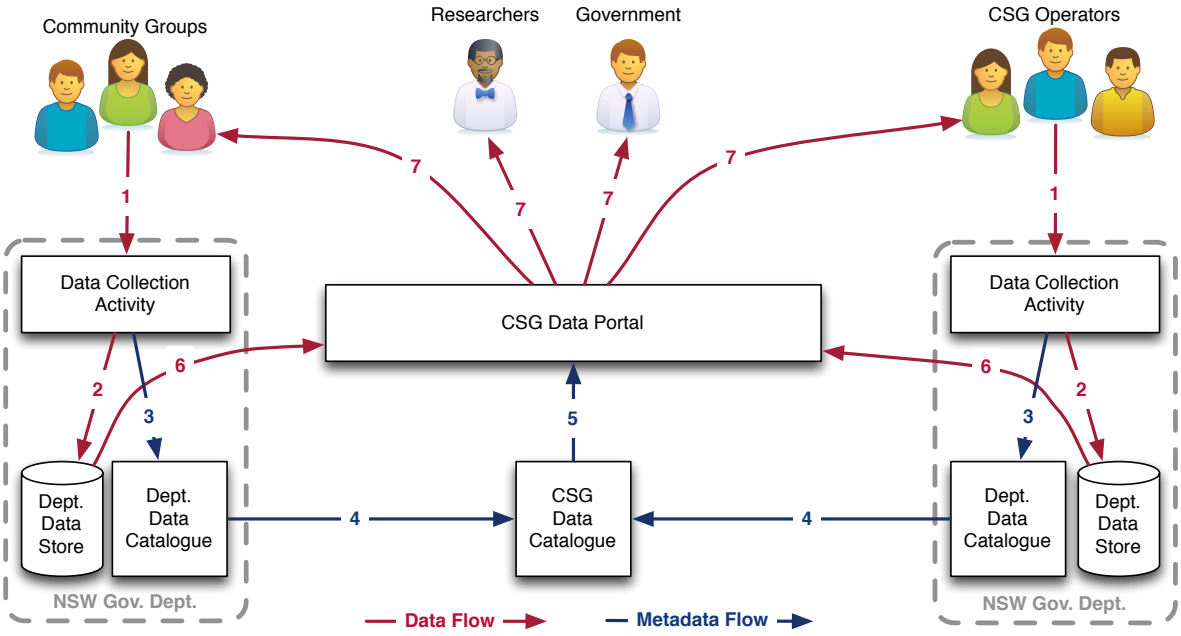
A federated approach has a number of advantages over other data management systems:

- Each data collecting entity has control over how they collect, manage and store their data. This is driven by the assumption that those collecting the data are the domain experts with the greatest understanding of the nuances of the data being collected;
- Storage and infrastructure costs for the federation can be distributed across the participating entities;
- Published data can be reused outside the original federation with relative ease;
- By making metadata available, existing datasets held by other entities can quickly be identified, reducing the collection of duplicate data;
- Access to live databases can be provided enabling real-time access to the latest data collected; and
- Each federation participant has control over the access permissions for their own datasets.

Federated data approaches do have some drawbacks. As the data within a data federation is the responsibility of individual federation participants, care must be taken to ensure that data curation and preservation activities are consistently applied across the federation. Additionally, the quality of data published in the federation may vary considerably depending on the data provider, as there is no centralised process for ingesting data. The quality of available data, however, should be clearly identified in the metadata describing the data allowing end-users of the data to evaluate its appropriateness.

5.1 Data collection and access

In Figure 1 below, the process of collecting CSG data in a federated approach is shown.



1. Data provided to various Government departments as part of their operational activities.
2. Data is stored and held within the department that has responsibility for collecting the data.
3. Metadata regarding the data is stored in the department data catalogue.
4. Each department publishes the metadata and this is aggregated into a CSG Data Catalogue.
5. The CSG Data Portal uses the CSG Data Catalogue to identify data available for different users.
6. The CSG Data Portal pulls data on-demand from the individual departments to make them available to the end-user.
7. All end-user to data is mediated through the CSG Data Portal.

Figure 1 – Example coal seam gas data federation

5.2 Technical Considerations

The construction of a data federation is a sizeable undertaking and requires careful coordination, especially during the initial stages. Critical to the success of federated data management and access systems is the development of the policies and processes surrounding the exchange of data and metadata. Some of the areas of consideration include:

- Metadata collection and specification;
- Dealing with confidentiality; and
- Identity, authentication and authorisation.

5.2.1 Metadata Collection and Specification

The metadata used to describe the data held by a federation is critical as it is the primary mechanism by which data is discovered and identified. Within the federation the metadata describing the data needs to be sufficient to allow users to identify that a particular dataset is appropriate for their needs. This requires consistent recording of metadata by each federation participant when collecting, sharing and publishing data within the federation.

There exist many metadata standards covering the full gamut of disciplines that collect data and it is often appropriate to use the metadata standard most appropriate for describing the type of data. This necessitates the use of a higher-level metadata description for describing the data independent of its type. To capture all the necessary high-level metadata for a CSG data federation it may be necessary to create a new, or tailor an existing, metadata standard.

5.2.2 Dealing with confidentiality

A federated system has a number of mechanisms for dealing with confidential data. First and foremost, each entity in the federation can decide which datasets to publish to the federation. Any data held that is confidential can simply not be made available in the federation. Similarly, when publishing datasets federation participants can choose to only make a subset of the data available, excluding any confidential information. Secure data environments are another approach to dealing with the issue of data confidentiality and privacy issues.

In a federated system it is the responsibility of the publishing entity to identify any security requirements, such as confidentiality and privacy, and to ensure these are encapsulated within the metadata description of the dataset.

5.2.3 Identity, Authentication and Authorisation

As has already been discussed, by its very nature a CSG data federation will contain data that is confidential and sensitive. This requires that users wishing to access data in the federation, where it is not openly available, are authorised to do so. There are two primary approaches to handling authorisation within a federation.

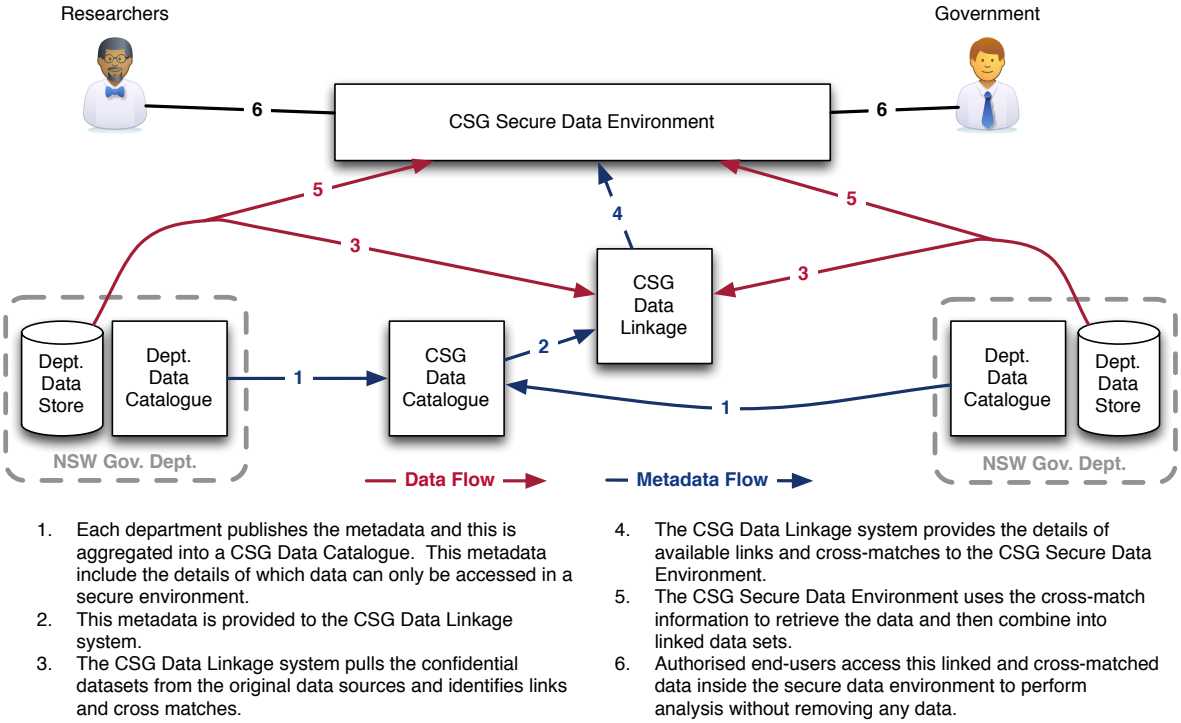
In the first model each federation participant handles the authentication and authorisation of users. This allows them to have very precise control over who is granted access. Unfortunately, this often results in significant duplication and confusion as users are required to register with each participant separately.

The second model is to have a central authority responsible for the identity management, authentication and authorisation of federation users. Having a central authority allows consistent processes to be applied in the registration, approval and categorisation of new users to the system. This in turn can simplify the authorisation of users to access individual datasets as federation-wide permissions can be granted to specific groups of users.

5.3 Secure environments

An advantage of the federated approach is that it allows secure environments to be created for the analysis of linked and cross-matched datasets. By leveraging the existing federation infrastructure a

secure data environment can be created. An example of how a secure data environment could be created for CSG data is shown below in Figure 2.



1. Each department publishes the metadata and this is aggregated into a CSG Data Catalogue. This metadata include the details of which data can only be accessed in a secure environment.
2. This metadata is provided to the CSG Data Linkage system.
3. The CSG Data Linkage system pulls the confidential datasets from the original data sources and identifies links and cross matches.
4. The CSG Data Linkage system provides the details of available links and cross-matches to the CSG Secure Data Environment.
5. The CSG Secure Data Environment uses the cross-match information to retrieve the data and then combine into linked data sets.
6. Authorised end-users access this linked and cross-matched data inside the secure data environment to perform analysis without removing any data.

Figure 2 – Example coal seam gas secure data environment

Appendix A NSW CSG Data Types and Sources

		New South Wales			QLD	Canada
	Type of Data	**EPA**	**Water**	**P&I**	**NR&M**	**AER**
Air	air emissions	**Y**				**Y**
	offensive odours	**Y**				
	climate conditions			**Y**		
Land	land characteristics			**Y**		
	local agricultural commodity production			**Y**		
	soil (types, depth, pH, salinity)			**Y**		
	topography/slope			**Y**		
Noise	noise	**Y**				
Water	bore standing water measurements				**Y**	
	groundwater - baseline conditions		**Y**		**Y**	
	groundwater flow model				**Y**	
	groundwater level		**Y**			
	groundwater pressure		**Y**			
	groundwater quality	**Y**	**Y**		**Y**	**Y**
	groundwater use quantity		**Y**			
	water characteristics (availability, quality)			**Y**		

		New South Wales			QLD	Canada
	Type of Data	**EPA**	**Water**	**P&I**	**NR&M**	**AER**
Operations	application submission, licence management	**Y**	**Y**	**Y**		**Y**
	gas removal data					**Y**
	gas well / infrastructure location	**Y**				
	incident reports					**Y**
	leak detection and repair records	**Y**				
	liability cost estimates					**Y**
	well drilling and completion data					**Y**
	well head configuration	**Y**				
	well packer test results					**Y**
	well test data					**Y**

Notes:

EPA NSW Environment Protection Authority

Water NSW Office of Water

P&I NSW Dept. of Planning & Infrastructure

NR&M QLD Dept. of Natural Resources & Mines

AER Alberta Energy Regulator (Canada)